IBM System z 10GbE RoCE Express feature with SMC-R (Shared Memory Communications – RDMA) VLAN Configuration Considerations

Sept 25, 2014

Jerry Stevens – sjerry@us.ibm.com

© 2014 IBM Corporation

SMC-R / RoCE VLAN Considerations

- ❑ Trunk Mode:
  When your OSA switch ports are configured in trunk mode, then your RoCE switch ports must also be configured in trunk mode and enabled for all associated OSA VLANs

- ❑ Access Mode:
  When your OSA switch ports are configured in access mode, then your RoCE switch ports must also be configured in access mode within a single VLAN

© 2014 IBM Corporation

Backup

Supporting background information and examples with additional details

1. Terminology (VLAN trunk and access modes)
2. Single IP Subnet examples:
   1. IP configuration
   2. RoCE configuration
   3. VLAN Configuration:
      1. Trunk Mode
      2. Access Mode
3. Multiple IP Subnet examples:
   1. IP configuration
   2. RoCE configuration
   3. VLAN configuration:
      1. Trunk mode
      2. Access mode
4. Summary (details)

© 2014 IBM Corporation

Terminology: VLAN Trunk / Access Mode Definitions

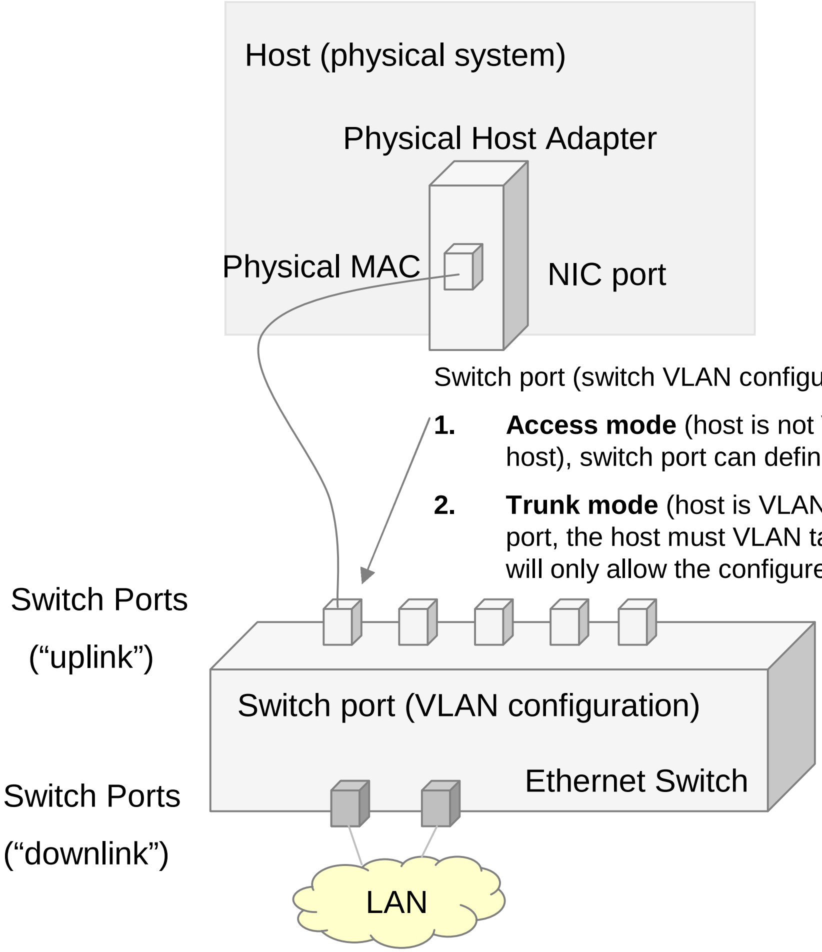
- Trunk Mode:
  Ethernet switch port configured to allow multiple VLAN IDs to / from the host adapter. The host OS is aware of VLAN IDs.
  - Outbound frames from the host contain (are tagged with) a VLAN ID by the host. The switch will allow frames that are tagged with valid VLAN IDs as defined by the switch trunk port.
  - Inbound frames from the switch to the host contain (are tagged with) a VLAN ID and filtered by the switch based on the switch port trunk definition.

- Access Mode:
  Ethernet switch port configured to allow a single VLAN ID to / from the host adapter. The host is unaware of VLAN IDs.
  - Outbound frames from the host do not contain (are not tagged) with a VLAN ID by the host. The switch will tag the frames with a single VLAN ID based on the switch port definition and forward into the LAN.
  - Inbound frames from the LAN are filtered by the switch. The switch will only allow a single VLAN ID from the LAN to this host port. The VLAN ID tag is removed before sending to the host.

Note. For additional details refer to IEEE 802.1Q.

© 2014 IBM Corporation

Ethernet Switch VLAN Terminology with Illustration



Host (physical system)

Physical Host Adapter

Physical MAC

NIC port

Switch port (switch VLAN configured mode):

1. **Access mode** (host is not VLAN aware, frames are not VLAN ID tagged to/from the host), switch port can define a single VLAN or no VLAN defined, switch tags and filters)
2. **Trunk mode** (host is VLAN aware, multiple VLAN IDs are configured in the switch port, the host must VLAN tag (outbound) and filter (inbound) packets). The switch will only allow the configured VLAN IDs via 802.1Q

Switch Ports

("uplink")

Switch port (VLAN configuration)

Ethernet Switch

Switch Ports

("downlink")

LAN

Native VLAN ID:

Typically considered a special

(zero) VLAN ID.

© 2014 IBM Corporation

RoCE, SMC-R and VLANs

- SMC-R connection eligibility requires that both host have access to the same IP subnet (i.e. RoCE is not routable).

- VLANs are optional for IP and RoCE connectivity

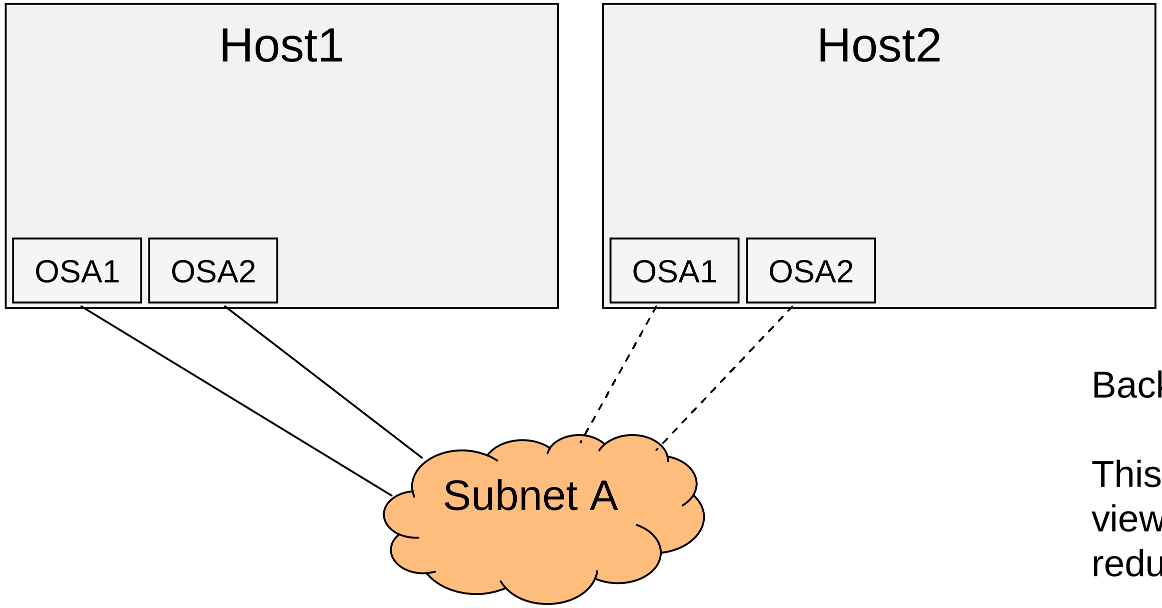
- When VLAN IDs are configured in the OS (OS is VLAN aware):
  - Indicates the switch port is configured in trunk mode
  - In z/OS VLAN IDs are configured in the TCP/IP profile on the OSA INTERFACE statement
  - VLAN IDs are dynamically propagated to RoCE for SMC-R Link Groups

- When VLAN IDs are not configured in the OS (OS is unaware of VLANs):
  - Indicates the switch port is configured in access mode
  - RoCE SMC-R Link Groups will not be VLAN qualified

- When VLANs are not in use for the IP connection (a variation of OS unaware) where untagged frames are allowed then the RoCE ports should also be untagged (or alternatively the RoCE ports must follow the access mode rules; use the same VLAN, transparent to the hosts).

© 2014 IBM Corporation

Example 1: Redundant IP configuration



Background:

This example illustrates a high level view of an IP configuration with redundancy.

Note that the two hosts have direct connectivity to the same IP subnet.

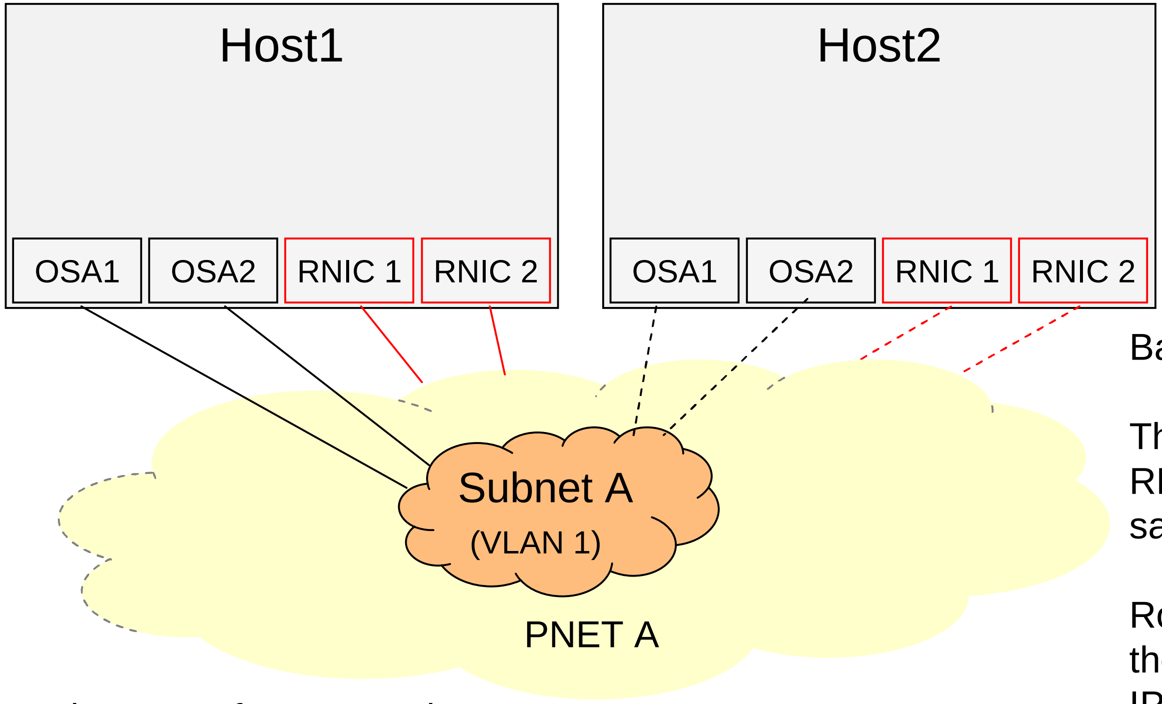
The redundant OSAs allows the host to use both paths and failover to a single OSA when one path becomes unavailable. Both static routing (using ARP takeover) and dynamic IP routing can be used for failover.

Since this is a single IP subnet the VLAN ID is not required to be exposed to the host (i.e. the VLANs could be configured in Access mode).

The VLANs could also be configured in trunk mode and therefore configured in the OS.

© 2014 IBM Corporation

Example 2: Redundant IP configuration with RoCE



The RoCE frames can be VLAN tagged by the host (trunk mode), the switch (access mode) or flow untagged.

When the SMC-R host is VLAN aware the RoCE traffic will be VLAN tagged by the host using the VLAN ID of the TCP connection.

Background:

This next example adds redundant RNICs for RoCE connectivity to the same physical network (“PNET A”).

RoCE frames do not flow over IP and therefore they are not associated with an IP subnet and are not IP routable. RoCE traffic must use direct L2 connectivity (VLANs are optional).

© 2014 IBM Corporation

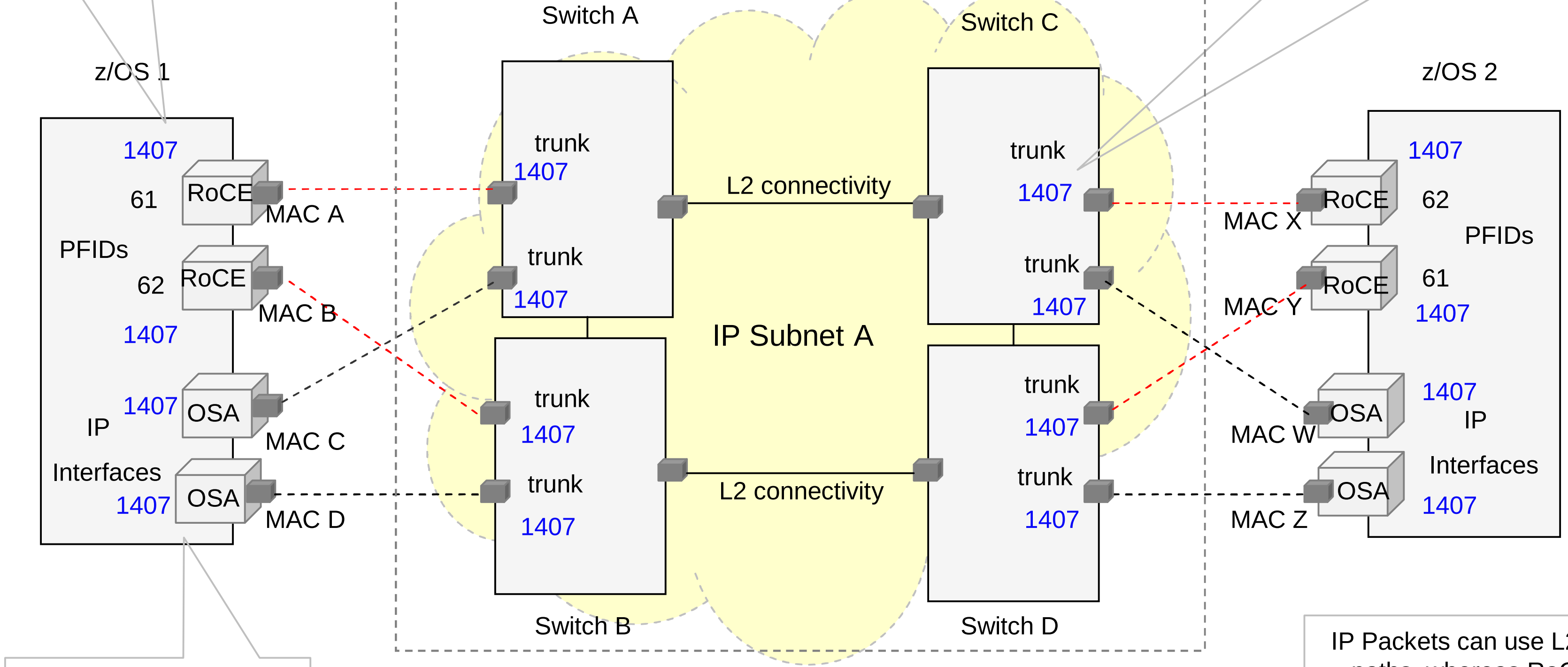
Example 3: VLAN Configuration (Trunk Mode)

All VLANs are configured in Trunk Mode

The RoCE VLANs are inherited from the OSA configuration and LGs (QPs) are VLAN qualified

Define all OSA and RoCE switch ports In trunk mode (RoCE ports must allow all valid VLAN IDs)

PNet A



When VLANs are configured in trunk mode they are visible to the host (configured on the OSA INTERFACE statement)

Possible SMC-R Link Group (e.g. RC-QPs for VLAN 1407)

QP 81 (MAC A) ◄-------------------► (MAC Y) QP 93

Link A

QP 82 (MAC B) ◄-------------------► (MAC X) QP 94

Link B

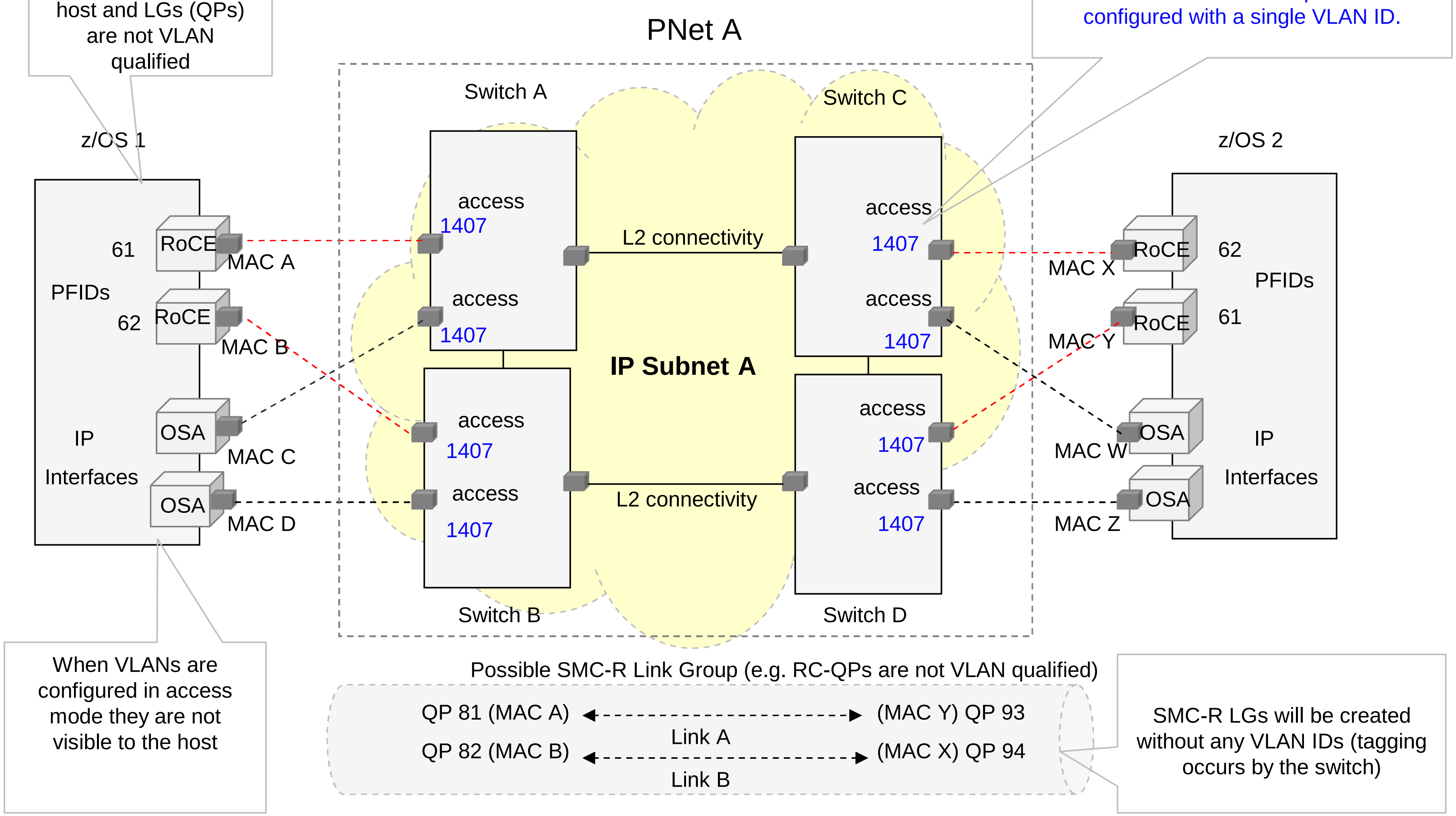
IP Packets can use L2 and L3 paths, whereas RoCE can only use the L2 path

SMC-R LGs will be created for each VLAN (e.g. 1407)

© 2014 IBM Corporation

Example 4: VLAN Configuration (Access Mode)

All VLANs are configured in Access Mode



The RoCE VLANs are transparent to the host and LGs (QPs) are not VLAN qualified

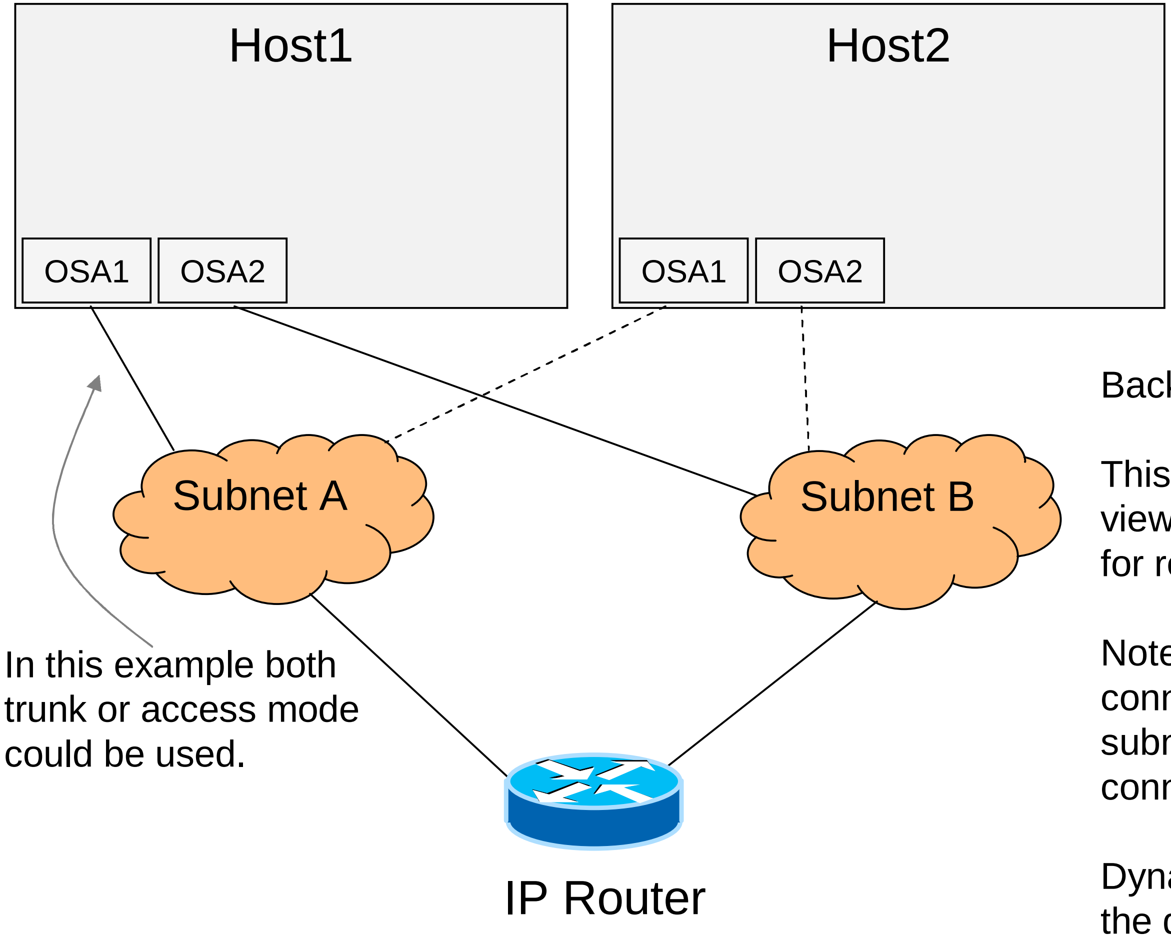
Define all OSA and RoCE switch ports in access mode. All RoCE ports must be configured with a single VLAN ID.

When VLANs are configured in access mode they are not visible to the host

SMC-R LGs will be created without any VLAN IDs (tagging occurs by the switch)

© 2014 IBM Corporation

Example 5: Redundant L2 / L3 IP configuration



In this example both trunk or access mode could be used.

Background:

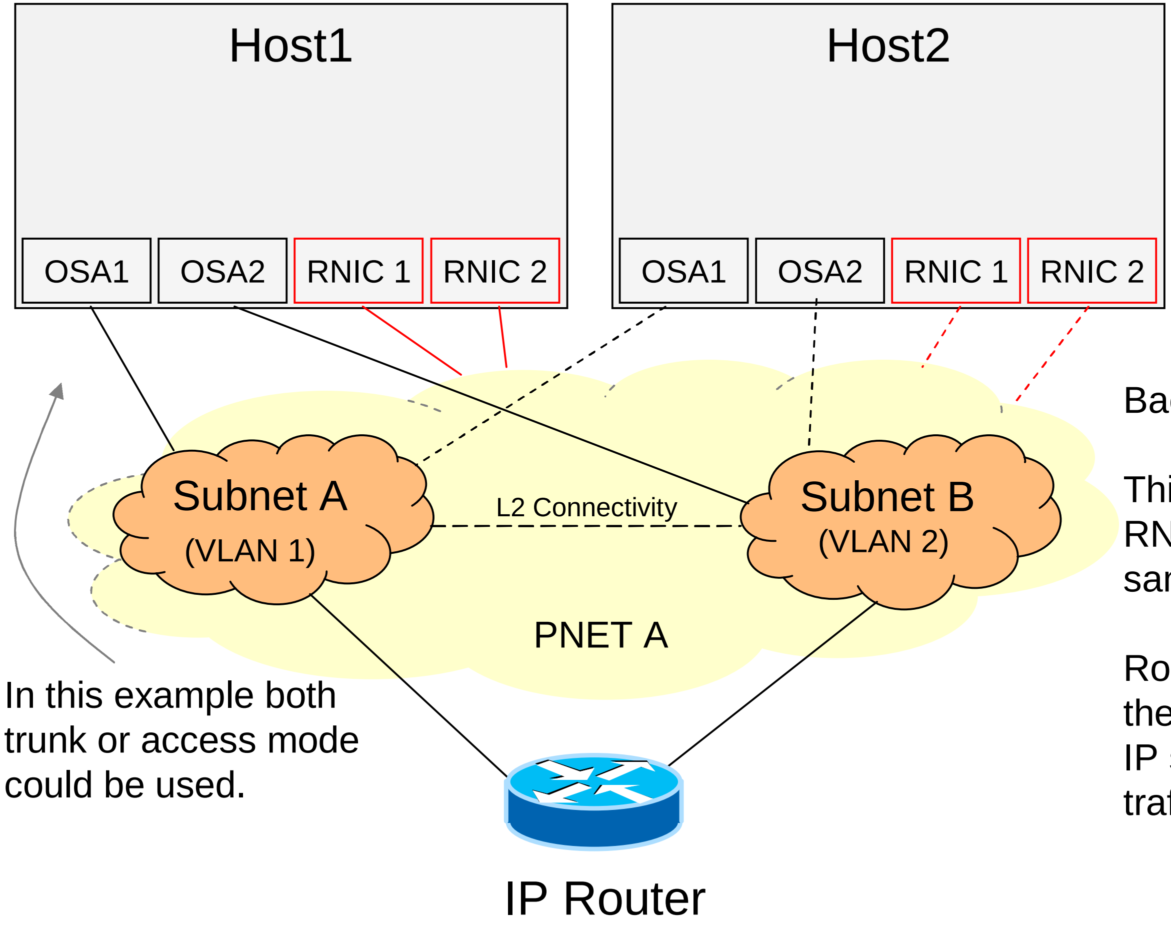
This example illustrates a high level view of a layer 2 / 3 IP configuration for redundancy using dynamic routes.

Note that the two hosts have direct L2 connectivity over two different IP subnets, and they also have routed L3 connectivity between the subnets.

Dynamic routing protocols will prefer the direct connectivity as long as it is available. When a single subnet is connected to all hosts then the connection is eligible for SMC-R.

© 2014 IBM Corporation

Example 6: Redundant IP configuration with RoCE



In this example both trunk or access mode could be used.

Background:

This next example adds redundant RNICs for RoCE connectivity to the same physical network (LAN “PNET A”).

RoCE frames do not flow over IP and therefore they are not associated with an IP subnet and are not IP routable. RoCE traffic must use direct L2 connectivity.

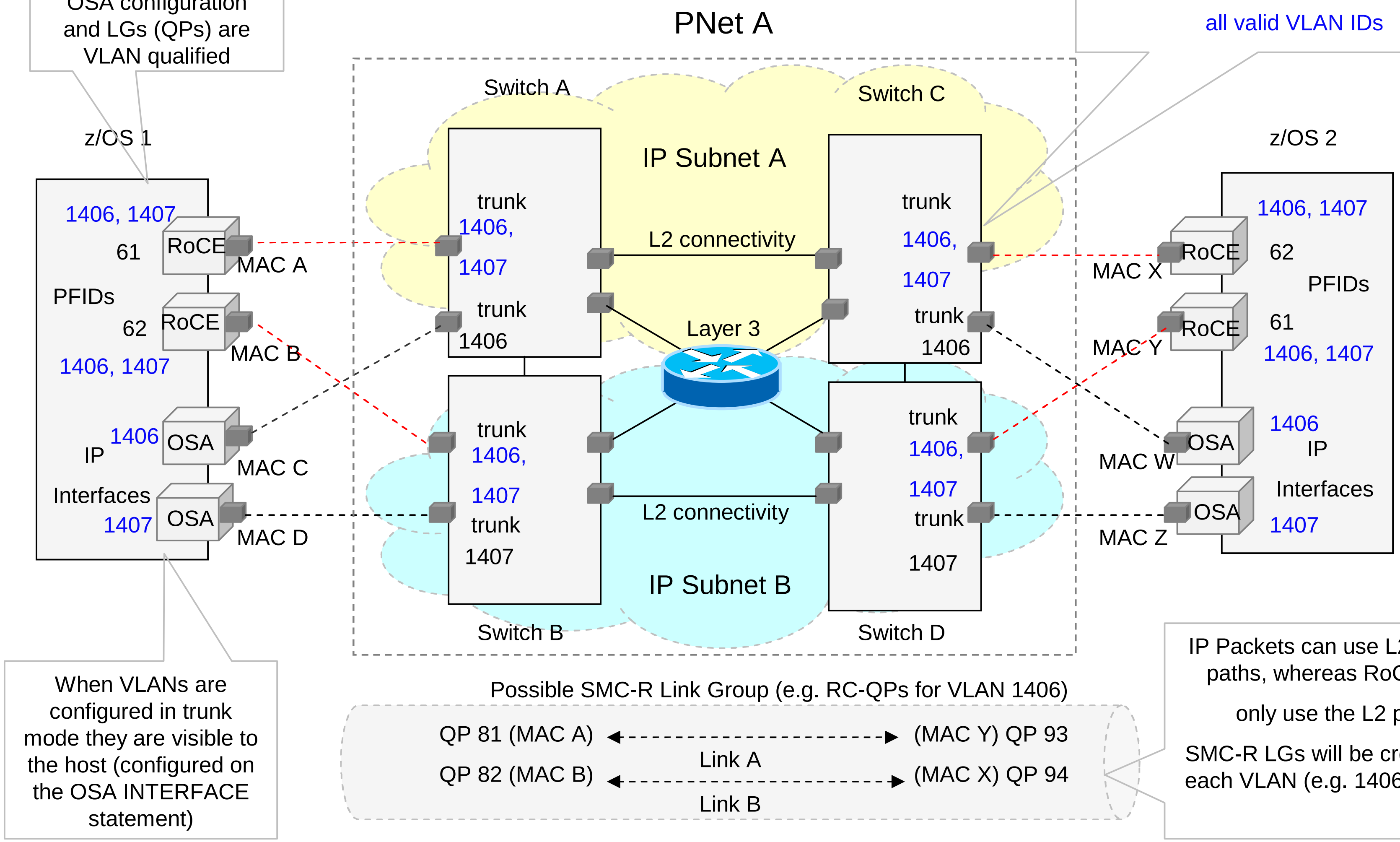
© 2014 IBM Corporation

Example 7: VLAN Configuration (Trunk Mode)

All VLANs are configured in Trunk Mode

The RoCE VLANs are inherited from the OSA configuration and LGs (QPs) are VLAN qualified

Define all OSA and RoCE switch ports in trunk mode. RoCE ports must allow all valid VLAN IDs



When VLANs are configured in trunk mode they are visible to the host (configured on the OSA INTERFACE statement)

IP Packets can use L2 and L3 paths, whereas RoCE can only use the L2 path

SMC-R LGs will be created for each VLAN (e.g. 1406 & 1407)

© 2014 IBM Corporation

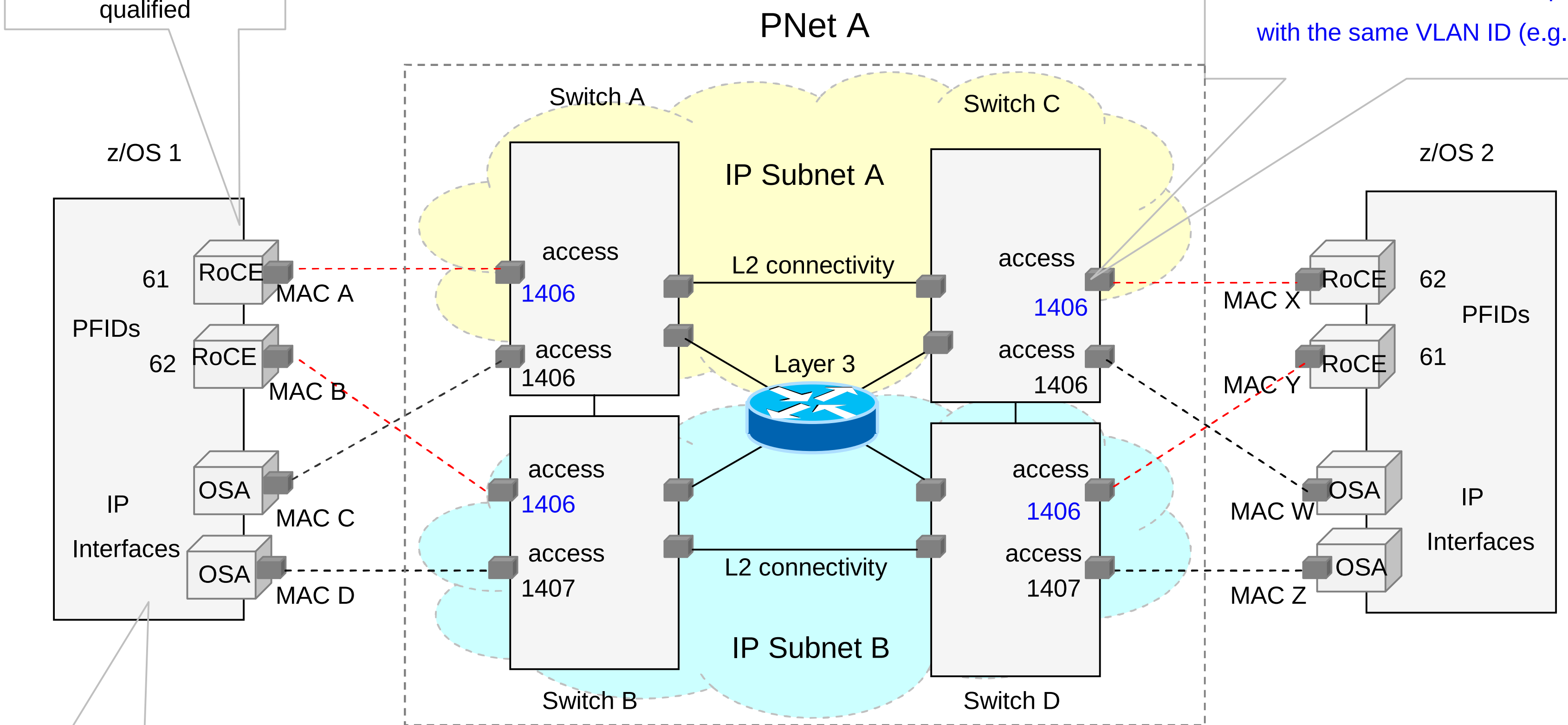
Example 8: VLAN Configuration (Access Mode Variation 1)

All VLANs are configured in Access Mode

The SMC-R LG created is not VLAN qualified

Access Mode Variation 1:

Define all RoCE switch ports with the same VLAN ID (e.g. 1406)



When VLANs are configured in access mode they are not visible to the hosts

Possible SMC-R Link Group (RC-QPs)

QP 81 (MAC A) ◄- - - - - - - - - - - - - - - - - - - ► (MAC Y) QP 93

Link A

QP 82 (MAC B) ◄- - - - - - - - - - - - - - - - - - - ► (MAC X) QP 94

Link B

Valid LG / Reason:

Now all 4 RoCE MAC addresses have connectivity (transparently on VLAN 1406)

© 2014 IBM Corporation

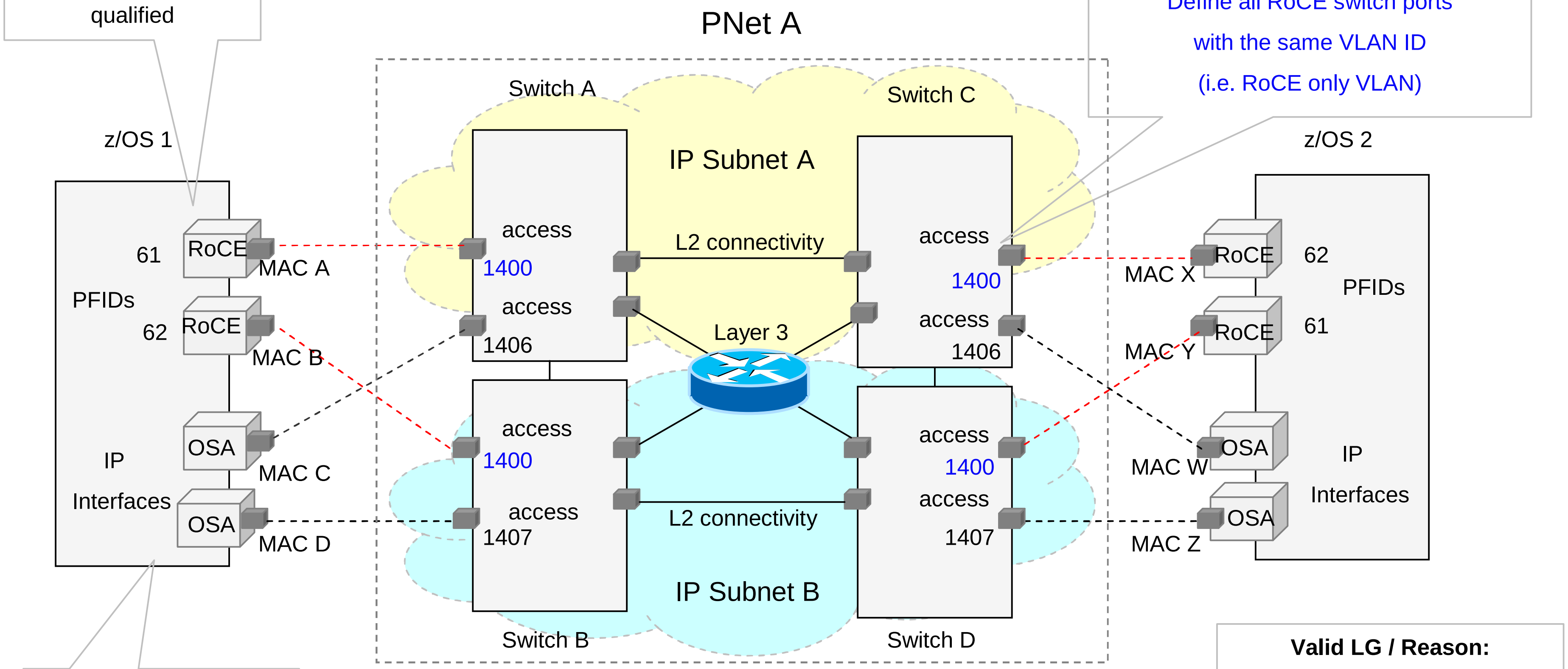
Example 9: VLAN Configuration (Access Mode Variation 2)

All VLANs are configured in Access Mode

The SMC-R LG created is not VLAN qualified

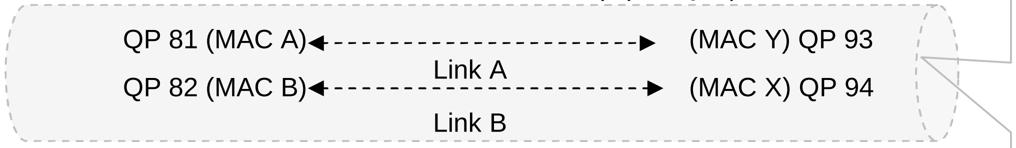
Access Mode Variation 2:

Define all RoCE switch ports with the same VLAN ID (i.e. RoCE only VLAN)



When VLANs are configured in access mode they are not visible to the hosts

Possible SMC-R Link Group (RC-QPs)



Valid LG / Reason:

Now all 4 RoCE MAC addresses have connectivity (transparently on VLAN 1400)

The RoCE VLAN ID does not have to match the OSA VLAN IDs in access mode

© 2014 IBM Corporation

Example 10: **Invalid** VLAN Configuration (Access Mode)

All VLANs are configured in Access Mode (but with multiple VLANs)

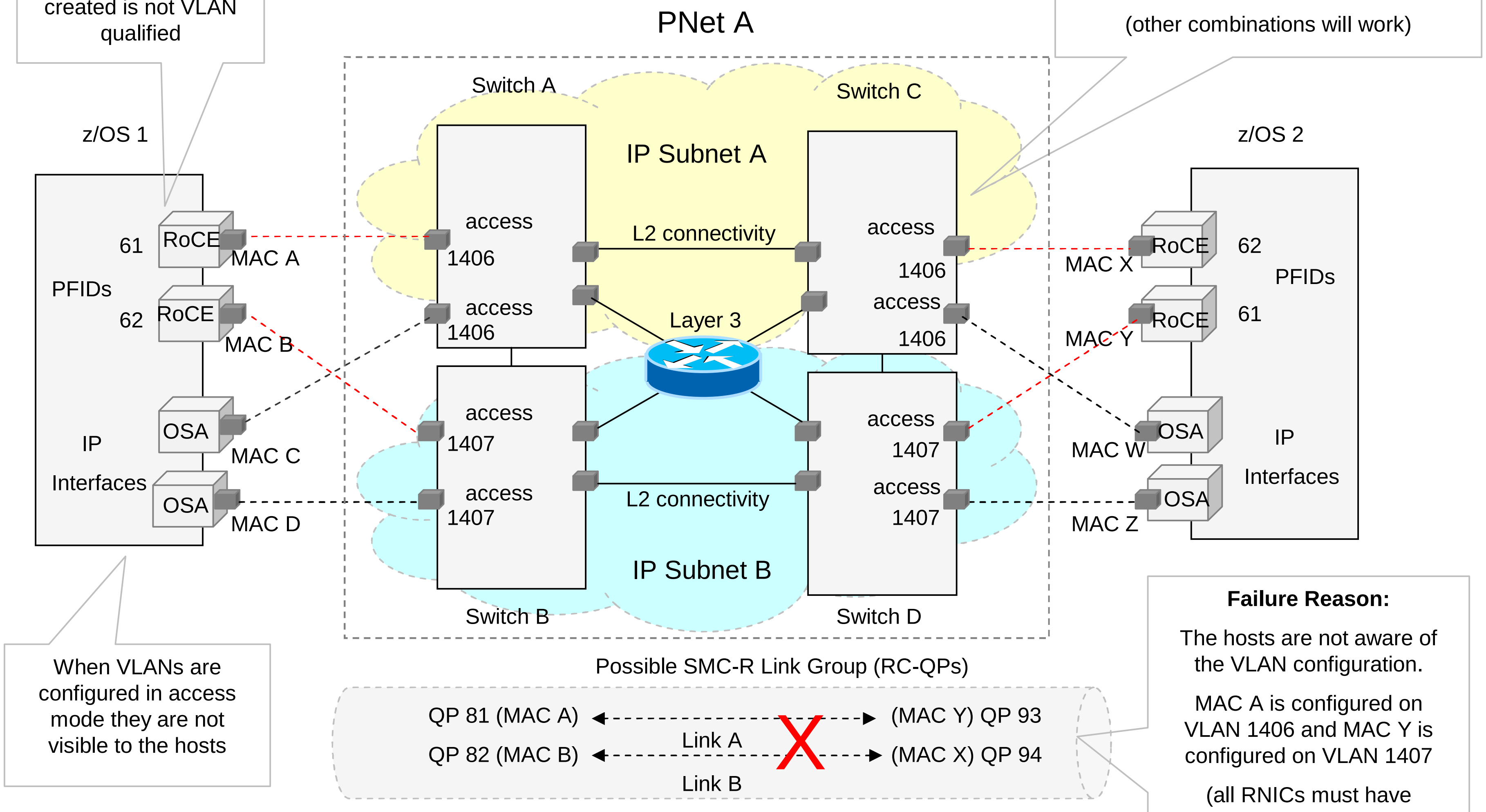
Invalid Configuration / Problem:

Packets from MAC A can’t reach MAC Y

Same is true for MAC B to MAC X

(other combinations will work)

The SMC-R LG created is not VLAN qualified



When VLANs are configured in access mode they are not visible to the hosts

Failure Reason:

The hosts are not aware of the VLAN configuration.

MAC A is configured on VLAN 1406 and MAC Y is configured on VLAN 1407

(all RNICs must have connectivity)

© 2014 IBM Corporation

Summary: SMC-R VLAN Configuration Rules

The following rules apply when using VLANs with SMC-R:

1. The Ethernet switch port VLAN mode must be consistent between the OSA Express Ethernet ports and their associated RoCE Express RDMA ports
   - If the OSA Express Ethernet switch ports are configured in trunk mode, their associated RoCE Express RDMA switch ports must also be configured in trunk mode
   - If the OSA Express Ethernet switch ports are configured in access mode, their associated RoCE Express RDMA switch ports must also be configured in access mode

2. The VLAN mode must be consistent between all of the hosts that will communicate over a LAN fabric (PNET) using SMC-R
   - You can't mix access and trunk modes among hosts on the same PNET if you are using SMC-R

3. The RoCE Express features must be on the same VLAN to communicate
   - If you are using **access mode**, the switch ports that are serving the RoCE Express features on a PNET must all be configured with the same VLAN ID. The RoCE VLAN ID is not required to match the VLAN ID of associated OSA Express features.
   - If you are using **trunk mode**, the RoCE Express features switch ports must be configured to allow the same VLAN IDs as the OSA Express features that they are associated with

© 2014 IBM Corporation